A Quasi-experimental Comparison of Econometric Models for Individual Health Care Expenditures

Partha Deb
Indiana University-Purdue University Indianapolis and
Hunter College, City University of New York

James F. Burgess, Jr.
US Department of Veterans Affairs Management Science Group and
Boston University School of Public Health

July 2002

Abstract

Individual health care expenditures have complex non-normal distributions with severe positive skewness and leptokurtosis. These features present severe challenges to reliable modeling of expenditures for prediction purposes. We compare a variety of methods using quasi-experimental techniques. Our quasi-experiments combine the distributional realism of actual data on health care expenditures with the robustness and reliability of Monte Carlo experiments. We find that models based on Gamma densities perform substantially better than models based on linear regression, with and without transformation. In addition, finite mixtures of Gamma densities offer some promising improvements over their one-component degenerate counterparts.

1 Introduction

Health care expenditures are non-negative random variables that can be statistically characterized by a non-trivial fraction of zero outcomes, a positively skewed empirical distribution of the positive values, and a density that is not easily characterized by known parametric forms. Jones (2001) discusses these and other analytical difficulties that plague research on health care expenditures differences across individuals. Our research, in which we focus on modeling positive health care expenditures, extends the literature in two important ways. First, we propose the use of finite mixture models for estimating health care expenditures which can serve as approximations to unknown probability densities (Lindsay, 1995; McLachlan and Peel, 1999). Second, we conduct an extensive evaluation of a number of econometric models in a quasi-experimental framework which combines the rigor of Monte Carlo experiments with the distributional realism of actual data.

Earlier work on modeling individual health care expenditures focused on the use of transformations of the dependent variable in linear regression models to improve the quality of estimates and predictions. Recent research has considered generalized linear models for estimating expenditures. Blough, Madden and Hornbrook (1999) demonstrate the feasibility of such models but do not directly compare their models with standard approaches. Manning and Mullahy (2001) show that the generalized linear model based on the Gamma density has promise, but also that some classes of generalized linear models are considerably more sensitive to data problems than OLS. In general, known, parametric densities are inadequate approximations to the true densities for health care expenditures, and robust estimators typically sacrifice precision.

Finite mixture models are, in principle, semiparametric and can approximate any probability densities. In practice, however, they should be viewed as flexible extensions of parametric models, potentially providing a compromise between strongly parametric and fully semiparametric models. Finite mixture models provide a natural and intuitively attractive representation of heterogeneity in a finite number of latent classes. The choice of the number of components in the mixture determines the number of classes, and the functional form for the density accommodates heterogeneity within each component. James Heckman's Nobel lecture states:

> "A major empirical finding in the work of Heckman and Singer that has been replicated in numerous subsequent studies is that distributions for unobservables can be approximated by low-dimensional finite mixtures or 'types'." (2001, p. 711).

Deb and Trivedi (1997, 2002) have demonstrated the superior performance of finite mixtures models for counts of health care utilization. Deb and Holmes (2001) show that a finite mixture model for positive mental health care expenditures provides more reliable estimates than the log regression model. Consequently, we evaluate a class of finite mixture models for health care expenditures in our quasi-experiment.

In addition to finite mixture models, we also consider linear regression models with and without transformations and generalized linear models (McCullagh and Nelder, 1989) based on the gamma density. We conduct an extensive evaluation of these models in a quasi-experimental framework. As in Monte Carlo experiments, confidence in results is achieved through replication. However, our experimental samples are not drawn from known distributions. Such data are unlikely to capture all the relevant features of the empirical distribution of health care expenditures. Instead, we assume that all relevant features of the empirical distribution of health care expenditures are present in the very large dataset we use so that sampling from it is equivalent to sampling from the distribution of health care expenditures in the population. To the extent that these data mimic features of health care expenditures in other populations, our quasi-experimental samples will be informative for models of health care expenditures in those populations.

We use annual patient expenditures and ICD-9-CM diagnoses from Fiscal Year 2000 US Department of Veterans Affairs (VA) as the basis for our quasi-Monte Carlo experiment. In the following section of the paper, we formally present the competing models used in this paper and discuss model comparison strategies. The data are described in section 3 and empirical results in section 4. We conclude in section 5.

2 Methods

2.1 Econometric Models

Let y_i denote health care expenditures for person i and x_i denote the set of covariates including the intercept. We estimate the following econometric models.

A linear conditional mean model is estimated using OLS so that

$$\begin{aligned} \widehat{\beta} &= \arg\min \sum_{i=1}^{N} \left\{ y_i - x_i\beta \right\}^2 , \qquad (1) \\ \widehat{y}_p &= x_p\widehat{\beta}, \end{aligned}$$

where $\widehat{y}_p$ denotes a conditional prediction. OLS with a linear mean has the desirable feature that it provides an unbiased predictor of health expenditures regardless of the distribution of the error term and the presence of heteroskedasticity. Nevertheless, given the extreme skewness of health care expenditures, it is possible that point forecasts obtained from this model may not be very precise. Note that this model is equivalent to the GLM model based on the normal density with linear link.

Two widely applied alternatives to the linear mean in the OLS context use transformations of the dependent variable. In the log model,

$$\begin{aligned} \widehat{\beta} &= \arg\min \sum_{i=1}^{N} \{\log(y_i) - x_i\beta\}^2 , \qquad (2) \\ \widehat{y}_p &= \exp(x_p\widehat{\beta}) \cdot \frac{1}{N} \sum_{i=1}^{N} \exp\left\{\log(y_i) - x_i\widehat{\beta}\right\} , \end{aligned}$$

and in the square root model,

$$\begin{aligned} \widehat{\beta} &= \arg\min \sum_{i=1}^{N} \{\sqrt{y_i} - x_i\beta\}^2 , \qquad (3) \\ \widehat{y}_p &= (x_p\widehat{\beta})^2 + \frac{1}{N} \sum_{i=1}^{N} \left\{\sqrt{y} - x_i\widehat{\beta}\right\}^2 , \end{aligned}$$

where the second term in each formula for the conditional prediction is a nonparametric smearing factor needed to retransform the prediction into the raw scale. Although these transformed models are designed to account for the skewness in health expenditures and the retransformation factors do not depend on normality of the errors, their predictions are not robust to heteroskedasticity in the transformed scale.

The model with the linear mean has an added advantage over models with complex mean specifications in that the regression coefficients are the average incremental costs of each disease and hence can be used to assess the face validity of the regressions. If used for rate setting, for example, plan managers would be very uncomfortable with negative regression coefficients or coefficient values outside the range of their intuitive expectations.

Often such models are recalibrated in ad hoc fashion until no "offending" coefficients remain (Ellis, R.P., personal communication). On the other hand, while the log and square root models generate positive conditional mean forecasts regardless of whether individual coefficients are positive or negative, the linear mean model may indeed generate negative predictions, which clearly lack face validity. To assess the consequences of imposing such face validity, i.e., restricting the conditional mean to be positive, we use the estimates from the linear OLS model to generate predictions of the form

$$\widehat{y}_p = \max(x_p\widehat{\beta}, 0). \tag{4}$$

The second set of models are in the GLM class. These models require only correct specification of the conditional mean for consistency and are quite flexible. We estimate GLMs based on the Gamma density as these have been shown to have desirable properties. We consider linear and square mean specifications so that

$$\begin{aligned} \widehat{\beta} &= \arg\max \sum_{i=1}^{N} \left\{ -\frac{y_i}{x_i\beta} + \log\left(\frac{1}{x_i\beta}\right) \right\}, \\ \widehat{y}_p &= x_p\widehat{\beta} \end{aligned} \tag{5}$$

and

$$\begin{aligned} \widehat{\beta} &= \arg\max \sum_{i=1}^{N} \left\{ -\frac{y_i}{(x_i\beta)^2} + \log\left(\frac{1}{(x_i\beta)^2}\right) \right\}, \\ \widehat{y}_p &= (x_p\widehat{\beta})^2, \end{aligned} \tag{6}$$

respectively.

Finally, we estimate 2 models that are based on finite mixtures of densities. The random variable y_i in a finite mixture model is assumed to be a drawn from an additive mixture of C distinct subpopulations or components in proportions $\pi_1, ..., \pi_C$, where $\sum_{j=1}^{C} \pi_j = 1$, $\pi_j \geqslant 0$ $(j = 1, ..., C)$. The mixture density for observation i, $i = 1, ..., n$, is given by

$$f(y_i|\boldsymbol{\theta}) = \sum_{j=1}^{C-1} \pi_j f_j(y_i|\boldsymbol{\theta}_j) + \pi_C f_C(y_i|\boldsymbol{\theta}_C), \quad i = 1, ..., n, \tag{7}$$

where $\pi_C = 1 - \sum_{j=1}^{C-1} \pi_j$. Each term in the sum on the right-hand side is the product of the mixing probability π_j and the component density $f_j(y_i|\boldsymbol{\theta}_j)$ which has parameters $\boldsymbol{\theta}_j$. In general, the π_j are unknown and estimated along with $\boldsymbol{\theta}_j$. A labelling restriction that $\pi_1 \geq \pi_2 \geq .... \geq \pi_C$, which can always be satisfied by rearrangement, is required for identification (normalization). Given our success with the gamma density in preliminary analysis, we consider models based on mixtures of gamma's:

$$\begin{aligned} \widehat{\beta}_j, \widehat{\pi}_j &= \arg\max \sum_{i=1}^{N} \log \left\{ \sum_{j=1}^{C-1} \pi_j \cdot \exp\left(-\frac{y_i}{x_i \beta_j} \right) \left(\frac{1}{x_i \beta_j} \right) \right\}, \qquad (8) \\ \widehat{y}_p &= \sum_{j=1}^{C-1} \widehat{\pi}_j x_f \widehat{\beta}_j, \quad j = 1, 2, ..., C. \end{aligned}$$

where β_j and π_j are estimated jointly.

We consider finite mixture models with linear mean specifications and two or three gamma component densities. Although both the specification of the mean and the number of components are trivially modified in principle, we restrict our attention to linear mean specifications for reasons of face validity discussed above and to two and three components for computational feasibility given the large scale of our study. Note that the model given by (8) is a generalization of (5), but it is possible that the two- and three-component mixture models perform worse than their one-component (degenerate) counterparts in finite samples.

Table 1 provides labels for each of the models considered in our experiments along with brief descriptions of the estimation method and prediction functions. The labels are subsequently used in our description of the results.

2.2 Experimental Design

The study design is quasi-experimental. Monte Carlo principles are used to create ‘experimental’ samples and confidence in results is achieved through experimental replication. However, unlike ‘true’ Monte Carlo experiments, our ‘experimental’ samples are drawn from real data with unknown distribution rather than artificial data drawn from a known distribution. It is well known that health care expenditures do not follow any known parametric distribution and that the characteristics of extreme observations make predicting health care expenditures a difficult exercise. Therefore, if we used data drawn from a known distribution in our study, it would likely not capture all the features of the empirical distribution of health care expenditures, and would have the additional drawback that it would always be possible to include an econometric model in the study that would a priori be closer to to the true data generating density (or even be correctly specified). Instead, we assume that all relevant features of the empirical distribution of health care expenditures are present in a very large dataset we use so that sampling from it is equivalent to sampling from the distribution of health care expenditures in the population. To the extent that these data mimic features of health care expenditures in other populations, our quasi-experimental samples will be informative for models of health care expenditures in those populations.

The dataset consisting of 2,979,760 observations was randomly split two groups: 1,500,000 observations were assigned to the estimation group and 1,000,000 to the prediction group. Note that these groups are quite large and reasonably might be treated as pseudo-populations. We were restricted to these sizes by computer memory considerations. Samples of size $N \in \{10000, 50000, 100000, 200000, 500000\}$ were drawn from the estimation group using simple random sampling with replacement. Note that most practical public or private populations in managed care plans or health care provider systems in the US fall in the range of our sample sizes (see e.g., Dunn, 1998, which analyzes risk adjustments in four populations of 240,000, 120,000, 115,000 and 70,000). The parameters of the models described above were estimated for each sample and saved. This process was repeated 20 times for each sample size. The parameters obtained from each replication were used to calculate conditional means using all million observations from the prediction group. Two statistics were calculated to evaluate the quality of the predictions: the mean prediction error

$$MPE = \frac{1}{N_f} \sum_{i=1}^{N_f} (\widehat{y}_f - y_i), \tag{9}$$

and the mean absolute prediction error

$$MAPE = \frac{1}{N_f} \sum_{i=1}^{N_f} |\widehat{y}_f - y_i|. \tag{10}$$

We also calculated each of these statistics after *trimming* the sample by eliminating 0.5% of the largest expenditures ($N_f = 995,000$) for two reasons. First, each of these statistics may be unduly affected by a very small fraction of extremely large expenditures in the prediction sample and these extreme observations may not regularly appear in smaller populations. Second, the design of many pricing schemes include reinsurance for very large expenditures so models should be evaluated on the observations not eligible for reinsurance.

Practitioners using models to determine costs of illnesses for rate-setting and other purposes sometimes top-code large expenditures to stabilize regression estimates. To evaluate the effects of such top-coding, we repeated the set of experiments described above with values of expenditures in the estimation samples *top-coded* at 100,000 (100,000 represents the 99.6^{th} percentile in the full dataset). We evaluated predictions based on these estimates using the prediction sample *unchanged* as well as *trimmed* and *top-coded*. We are agnostic regarding the merits of top-coding and trimming. Thus, we present results for all sets of experiments below.

2.3 Response Surfaces

For any statistic of interest, ideally one would like to compute analytical formulae for its predicted values as a function of experimental characteristics. For example, if the statistic of interest is bias and one is interested in determining how bias decreases as the sample size increases, the ideal would be an analytical formula that related bias to sample size. If these are not known, it is possible to approximate them using polynomial approximations to the true functional forms. Regressions of these polynomial approximations are called response surfaces. Response surface methodology facilitates understanding of experimental evidence because large amounts of experimental data can be summarized using simple functional forms. It also provides applied researchers a simple tool for computing outcomes at points in the design space that are not included in the experimental study. Another advantage, especially for computationally intensive processes, is that a large number of replications is not required. See Maasoumi and Phillips (1982), Hendry (1982), and Davidson and MacKinnon (1993) for detailed discussions of the merits of response surface methodology. Each of these advantages of response surface methodology is important in the context of our study relative to simple tabulation of the results: we have many design points (model×sample size), interest in performance at other sample sizes, and very computationally intensive estimation.

Let the models in this study be numbered by $m = 1, 2, ..., 8$. Let $j = 1, 2, ..., (8 \times 5 \times 20)$ denote an observation consisting of the statistics of interest (MPE_j, $MAPE_j$) and $d[m]_j$ which are dummy variables indicating the model on the basis of which the statistic was calculated. Let N_j denote the sample size used for estimation of the model. The response surfaces for MPE is specified as

$$MPE_j = \sum_{m=1}^{8} \alpha[m]d[m]_j + \sum_{m=1}^{8} \frac{\gamma[m]d[m]_j}{N_j} + u_j, \tag{11}$$

where $\alpha[m]$ and $\gamma[m]$ are regression coefficients. The second term in the right hand side of the regression reflects the fact that MPE is expected to decline at the rate N. Note that in each response surface regression, $\alpha[m]$ denotes the asymptotic expected value of MPE for model m. Expected MPE for desired finite sample sizes can be calculated by plugging in those sample sizes.

$MAPE$ only takes positive values, so its response surface is specified in logarithms, i.e.,

$$\log(MAPE_j) = \sum_{m=1}^{8} \alpha[m]d[m]_j + \sum_{m=1}^{8} \frac{\gamma[m]d[m]_j}{N_j} + u_j. \tag{12}$$

Now differences in values of $\alpha[m]$ represent percentage differences in $MAPE$ across models.

3 Data

The VA operates the largest health care system in the US with 163 hospitals, more than 800 community and facility-based clinics, 135 nursing homes, and other facilities. With a medical care budget of more than $19 billion in

FY2000, VA provided care to 3.8 million unique users, 3,000,499 of whom were provided care under priority for service connected disabilities, meeting an income/wealth based means test, or from a variety of smaller health care need and veteran specific reasons. 2,979,760 of these patients have measured costs accurate enough to be included in the patient sample that serves as the sampling population for our analysis as described above.

In recent years, the most important advances in risk adjusting patient populations to explain health care expenditures have employed diagnostic information to characterize disease patterns. There are two basic strands of analysis that flow from this work. Most commonly, analysis has focused on predicting the health care utilization of enrolled patient populations next year from diagnoses and other information (possibly even including costs) collected this year. This is called prospective modeling. However, a growing application of risk adjustment is in helping integrated health care delivery systems or insurers understand differences in the risk of current populations, for budget allocation or rate setting purposes. We employ this type of concurrent modeling in this paper.

We provide summary statistics for costs in Table 2. The estimates are based on a sample size of 2,500,000 that comprise the combination of our estimation and prediction samples. As is well known for health care expenditures in other contexts, expenditures for the VA population are also highly skewed and leptokurtic. When logarithms of health care costs are examined, skewness and kurtosis are considerably smaller but statistically significant. As a comparison, we also report summary statistics of health care expenditures for a representative sample of the US population in 1996 obtained from the Medical Expenditure Panel Survey (MEPS) and for the sub-sample of MEPS respondents enrolled in Medicare. The results in Table 2 show that the statistical characteristics of health care expenditures of the Medicare population are very similar to those of the VA population and that the distribution of health care expenditures for the US population overall are considerably more skewed and leptokurtic than either of the sub-populations. Note that as the data are refined into more homogeneous populations, the skewness and kurtosis moment measures fall.

To characterize the explainable portion of variation in expenditures, we employ Diagnostic Cost Group (DCG)/Hierarchical Coexisting Conditions (HCC) models (Ellis, et al. (1996), Ash, et al. (1998), Pope, et al. (1998)) to group ICD-9-CM diagnoses into HCC indicator groups as explanatory variables for health care expenditures. This model takes the 15,000 ICD-9-CM codes, groups them into categories and then places the groups into body system/clinical condition specific hierarchies. These hierarchies allow some multiple HCC's and disallow others, helping to address overfitting problems when people with complex diagnoses also by definition have less complex ones in the same hierarchy. Out of the 118 HCC's in Version 5 of the DCG model, we employ 42 HCC's in our model that appear with a frequency of at least 1 percent in the sample of 2,500,000. Brief descriptions of the HCC's and their sample frequencies are reported in Table A1.

4 Results

As described above, 5 different sample sizes were considered for estimation and each experiment was replicated 20 times. OLS estimates are trivially obtained. The log likelihood functions of GLM models with gamma baseline density are typically well behaved so ML estimation is easy to conduct, though obviously computationally intensive for some of the larger sample sizes. The log likelihood functions of finite mixture models are not so well behaved in principle. They can have multiple optima. In practice one can overcome this potential problem simply by experimentation with starting values although more complex algorithms are also available to avoid convergence to local optima. But in this experimental setting, it was not feasible to ensure convergence to the global maximum in each case. For the two-component mixture model, we used starting values based on the converged estimates of the Gamma model (degenerate mixture) which it generalized. For the three-component mixture model, we used starting values based on the converged estimates of the two-component mixture. Although these are reasonable starting values, convergence to a local optimum cannot be ruled out. Therefore, the results for the finite mixture models may be contaminated by non-maximized estimates, thus should be treated as the worst case scenarios.

The experimental samples of MPE and $MAPE$ consist of 800 observations each. Response surfaces regressions specified as (11) and (12) samples were estimated and the parameter estimates are reported in Tables 3 and 4. In each case, the R^2 of the regressions are over 0.99 demonstrating that they are very well specified and capture most of the variation across design points.

The asymptotic expected values of MPE indicate how the average value of predicted health care expenditure from a particular model compares to the average health care expenditure in the prediction sample. The results in Table 3 show that the linear and square root regression models estimated by OLS have negligible bias when the sample used to evaluate the predictions is created under the same rules as the sample used to estimate the model, i.e., when both estimation and prediction samples are either *unchanged* from the raw data or are *topcoded*. All other models have substantially larger biases. Predictions from both finite mixture models are downward biased. However, FM3-Γ-linear has lower bias than FM2-Γ-linear. In the *trimmed* prediction samples, predictions from the finite mixture models [FM2-Γ-linear and FM3-Γ-linear] have the smallest biases. Both ols-linear and ols-square root are upward biased. The results for the *unchanged* and *trimmed* prediction samples taken together indicate that the lower bias of the linear OLS model is due to its ability to predict the largest expenditures well. Note also that the OLS model with non-negative predictions [ols-linear>0] has a significant bias in each case, overpredicting relative to the standard OLS model [ols-linear] by about $60 on average.

As discussed above, the choice of functional form for the specification of the conditional mean is important for a variety of reasons. Although the linear conditional mean has virtues in its simplicity and ease of interpretation, the square and exponential conditional means have other attractive virtues. The results show that the log regression model performs surprisingly poorly vis-a-vis the alternatives. It produces substantially upward biased predictions. In preliminary work we found that the gamma model with exponential link also performed very poorly hence was eliminated from further consideration. We have chosen to include the log regression model because it is a leading model among those used in existing empirical studies. Within the family of linear regression models, the linear and square root models have very similar MPE's. It is not possible to discriminate between the two models on this basis. In the case of GLM models based on the Gamma density, there are differences in MPE's between linear and square links, but neither dominates.

The MPE of FM2-Γ-linear is smaller than the MPE of FM3-Γ-linear in every case. Given that FM2-Γ-linear generally underpredicts expenditures, these features suggest that the third component in FM3-Γ-linear captures the heterogeneity inherent in some of the very large expenditures. It is plausible that additional components would provide further improvement in MPE.

The asymptotic expected values of $\log(MAPE)$ indicate how values of individual predicted health care expenditures from a particular model compare to the values of actual health care expenditures in the prediction sample. Models with lower values of $\log(MAPE)$ predict individual expenditures better than models with higher values. The results in Table 4 show that the two-component finite mixture model with gamma densities dominates the rest by the $MAPE$ criterion. When both estimation and prediction samples are either unchanged from the raw data, FM2-Γ-linear has an 11 percentage point lower $MAPE$ than the linear regression model and 2 percentage point lower $MAPE$ than Γ-linear. Interestingly, the $MAPE$ of FM3-Γ-linear is worse than FM2-Γ-linear and, indeed, is comparable to the $MAPE$ from Γ-linear. There are two potential reasons for this decline in performance. First, it is possible that this is manifestation of a $MPE - MAPE$ trade-off which appears is many statistical contexts. In our context, it would most likely be due to the fact that by increasing the ability of the model to predict the small number of high expenditures well, the three-component model was doing worse at predicting the large number of lower expenditures. Second, it is possible that the parameter estimates of three-component model are based, in a substantial fraction of cases, on log likelihood values that are not globally maximized. This is plausible because the log likelihood function of the three-component model have multiple optima, in principle. Unfortunately, a closer investigation of these possibilities is beyond the scope of this paper.

The OLS model with logarithmic transformation continues to perform very poorly. The $MAPE$ from the square root model is always lower than the $MAPE$ from the linear model in the regression case, but the relative performance of the linear and square conditional means are reversed in the Γ GLM models.

The finite sample values of MPE and $MAPE$ and their rates of convergence to the asymptotic values of are also described by the estimates in Tables 3 and 4 respectively, but these are obviously not transparent. Therefore, in Figures 1 and 2, we plot the values of MPE and $\log(MAPE)$ expected at different sample sizes for three of the leading models with linear mean specifications - ols-linear, Γ-linear and FM2-Γ-linear. As was evident from the regression estimates, ols-linear has the lowest bias when prediction samples are unchanged, but that models in the gamma family, Γ-linear in one case and FM2-Γ-linear in the other, have lower biases when the prediction sample is trimmed. The rates of convergence of ols-linear and Γ-linear to the asymptotic MPE is very quick; 15,000-20,000 observations appear to be sufficient. On the other hand, convergence is slower for the mixture model, as expected. But even for FM2-Γ-linear sample sizes of 30,000-40,000 are sufficient to ensure asymptotic values of MPE.

$MAPE$ converges at similar rates. The advantages of the models based on the gamma density vis-a-vis ols-linear are dramatic. The gains from using FM2-Γ-linear are in the order of 2-3 percentage points for sample sizes over 20,000. In the context of the budgets at stake in many rate-setting exercises, these gains are substantive.

5 Conclusion

Many health outcome variables in health economics deviate from known parametric densities even upon tranformation and reliable estimation methods for practical purposes continues to be an unsettled issue. The pseudo-Monte Carlo experiments reported in this paper subject a number of plausible econometric models to tests in a variety of dimensions. The results demonstrate that models with linear mean specifications perform at least as well as models with more complex means or those that require retransformation. Linear regression models estimated by ordinary least squares produce unbiased predictions, but individual predictions relatively imprecise. Because there exist incentives for providers to miscode, misreport, etc. when payments deviate substantially from costs, unbiased predictions are an inadequate criterion for a good econometric model. The ideal standard involves predicting individual expenditures as well as possible given the data. A GLM model based on the Gamma density with a linear link has reasonable bias properties and superior individual predictions vis-a-vis the linear regression model. A finite mixture model constructed using two Gamma densities with linear means has lower biases in some cases and superior individual predictions in every case relative to the GLM model based on the Gamma density. Adding a third component to the mixture model appears to improve biases but at the cost of poorer individual predictions.

In practice, estimation of finite mixture models raises some computational difficulties, especially as the number of points of support in the mixture distribution increases. This paper shows that there are substantial gains in predictive performance associated with the use of finite mixture models with two components, The finding is consistent with conventional wisdom and empirical evidence in the literature on finite mixture models that two to four points of support are typically sufficient. A small number of components is more likely to be sufficient if one starts with a baseline density that forms a reasonable first approximation to the true data density. Therefore, given the advances in computer hardware and statistical computing technology, the computational burden of finite mixture models should not discourage its use.

References

Ash A, R. P. Ellis, W. Yu, et al. (1998), "Risk Adjustment for the Non-Elderly." Final Report submitted to the Health Care Financing Administration under cooperative agreement No.18-C-90462/1-02, Boston, MA: Boston University, June 1998.

Blough, D.K., C.W. Madden and M.C. Hornbrook (1999), "Modeling Risk using Generalized Linear Models, *Journal of Health Economics*, 18, 153-171.

Davidson, R. and J. G. MacKinnon (1993), *Estimation and Inference in Econometrics*, Oxford University Press.

Deb, P. and A. M. Holmes (2000), "Estimates of Use and Costs of Behavioral Health Care: A Comparison of Standard and Finite Mixture Models", *Health Economics*, 9, 475-489.

Deb, P., and P.K. Trivedi (1997), "Demand for Medical Care by the Elderly in the United States: A Finite Mixture Approach", *Journal of Applied Econometrics*, 12, 313-336.

Deb, P., and P.K. Trivedi (2002), "The Structure of Demand for Health Care: Latent Class versus Two-part Models", *Journal of Health Economics*, 21, 601-625.

Dunn, D. (1998), "Applications of Health Risk Adjustments: What Can Be Learned to Date?" *Inquiry*, 35: 132-147, Summer.

Ellis, R.P., G. Pope, L.I. Iezzoni, J.Z. Ayanian, D.W. Bates, H. Burstin, and A. Ash (1996), "Diagnosis-Based Risk Adjustment for Medicare Capitation Payments." *Health Care Financing Review*, Spring.

Heckman, J.J. (2001), "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture", *Journal of Political Economy*, 109, 673-748.

Hendry, D. F. (1982), "A Reply to Professors Maasoumi and Phillips", *Journal of Econometrics*, 19, 203-213.

Jones, A. M. (1999), "Health Econometrics", forthcoming in J. P. Newhouse and A. J. Culyer, eds., *Handbook of Health Economics*, Amsterdam: North - Holland.

Lindsay, B. J. (1995), *Mixture Models: Theory, Geometry, and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5, IMS-ASA.

Maasoumi, E. and P. C. B. Phillips (1982), "On the Behavior of Inconsistent Instrumental Variable Estimators", *Journal of Econometrics*, 19, 183-201.

Manning, W.G. and J. Mullahy (2001), "Estimating Log Models: To Transform or Not to Transform?", *Journal of Health Economics*, 20, 461-494.

McCullagh, P. and J.A Nelder (1989), *Generalized Linear Models*, London: Chapman and Hall.

McLachlan, G.J., and D. Peel (2000), *Finite Mixture Models,* New York: John Wiley.

Pope, G., R.P. Ellis, C. Liu, et al. (1998), "Revised Diagnostic Cost Group (DCG)/Hierarchical Coexisting Conditions (HCC) Models for

Medicare Risk Adjustment." Final Report to the Health Care Financing Administration under Contract No. 500-95-048, Waltham, MA: Health Economics Research, Inc., February 1998.

Table 1
Description of Models

	Label	Estimation method	Prediction function
1	ols-linear	OLS	$x_i\beta$
2	ols-log	OLS	$\exp(x_p\widehat{\beta}) \cdot \frac{1}{N}\sum_{i=1}^{N} \exp\left\{\log(y_i) - x_i\widehat{\beta}\right\}$
3	ols-square root	OLS	$(x_p\widehat{\beta})^2 + \frac{1}{N}\sum_{i=1}^{N}\left\{\sqrt{y} - x_i\widehat{\beta}\right\}^2$
4	ols-linear>0	OLS	$\max(x_i\beta_j, \varepsilon)$
5	Γ-linear	ML, Γ density	$x_i\beta$
6	Γ-square	ML, Γ density	$(x_i\beta)^2$
7	FM2-Γ-linear	ML, mixture of 2 Γ's	$\sum_{j=1}^{C-1}\widehat{\pi}_j x_f \widehat{\beta}_j, \quad j = 1, 2$
8	FM3-Γ-linear	ML, mixture of 3 Γ's	$\sum_{j=1}^{C-1}\widehat{\pi}_j x_f \widehat{\beta}_j, \quad j = 1, 2, 3$

Table 2
Summary Statistics of Costs

	Cost/$1000			log(Cost/$1000)		
	VA	MEPS		VA	MEPS	
		All	Medicare		All	Medicare
N	2500000	18490	2588	2500000	18490	2588
Mean	5.342	2.372	6.185	0.411	-0.564	0.738
Median	1.537	0.527	2.097	0.430	-0.641	0.741
Std Deviation	14.804	8.572	12.521	-0.102	1.637	1.549
Skewness	9.717	21.287	6.388	-0.102	0.212	-0.201
Kurtosis	203.512	850.131	68.772	0.697	-0.151	0.161
99th percentile	70.322	29.852	58.440	4.253	3.396	4.068
95th percentile	22.612	9.491	25.773	3.118	2.250	3.249
75th percentile	3.839	1.699	5.883	1.345	0.530	1.772
25th percentile	0.586	0.180	0.798	-0.534	-1.715	-0.225
5th percentile	0.107	0.180	0.149	-2.235	-3.147	-1.904
1st percentile	0.032	0.180	0.038	-3.442	-4.605	-3.270

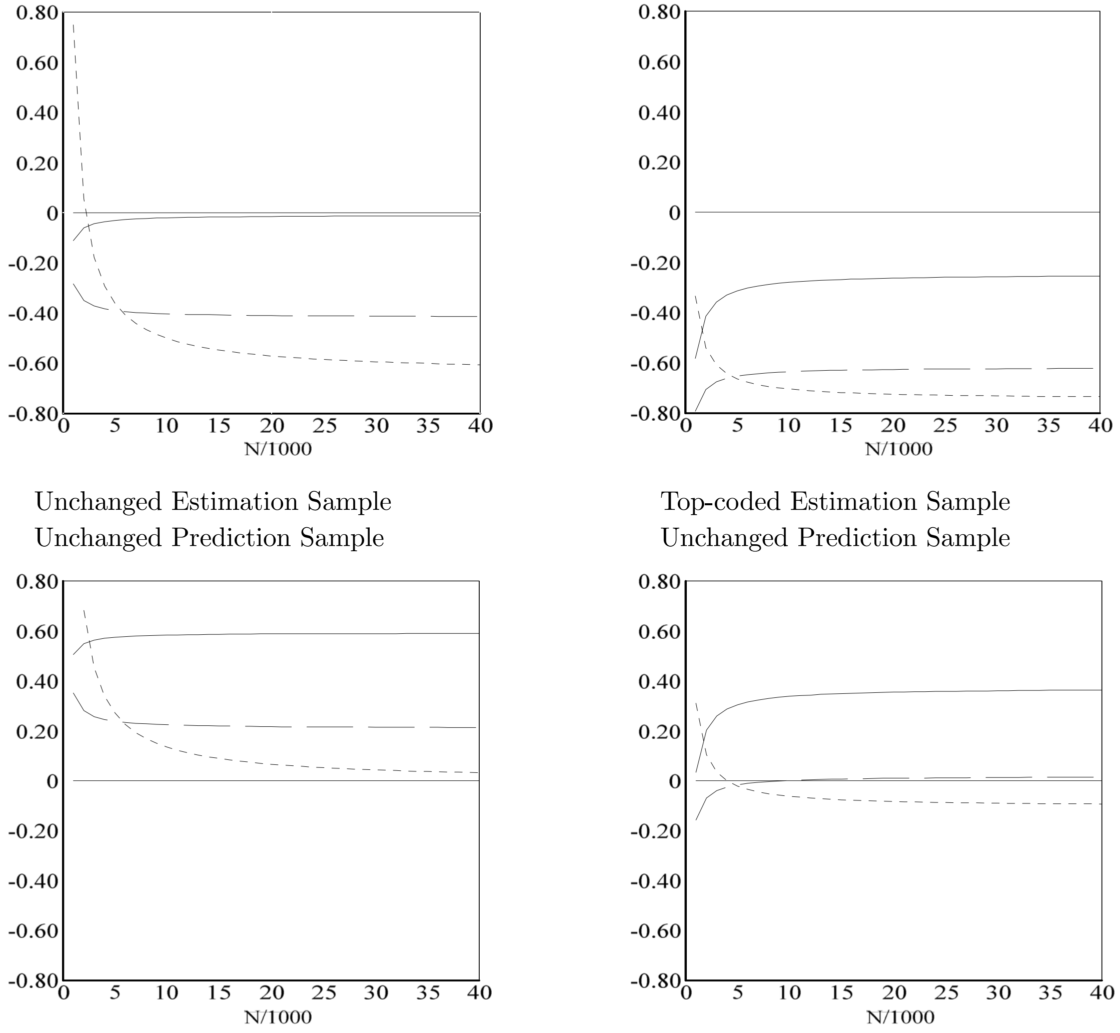
Table 3
Response Surface Regressions for Mean Prediction Error

Estimation Sample	Unchanged	Unchanged	Top-coded	Top-coded	Top-coded
Prediction Sample	Unchanged	Trimmed	Unchanged	Trimmed	Top-coded
ols-linear	-0.010	0.592	-0.246	0.372	-0.002
	(0.014)	(0.013)	(0.012)	(0.011)	(0.012)
ols-log	3.932	3.965	3.589	3.667	3.833
	(0.014)	(0.013)	(0.012)	(0.011)	(0.012)
ols-square root	-0.013	0.604	-0.248	0.376	-0.004
	(0.014)	(0.013)	(0.012)	(0.011)	(0.012)
ols-linear>0	0.050	0.653	-0.194	0.423	0.049
	(0.014)	(0.013)	(0.012)	(0.011)	(0.012)
Γ-linear	-0.417	0.210	-0.617	0.019	-0.374
	(0.014)	(0.013)	(0.012)	(0.011)	(0.012)
Γ-square	0.138	0.716	-0.082	0.510	0.161
	(0.014)	(0.013)	(0.012)	(0.011)	(0.012)
FM2-Γ-linear	-0.641	-0.002	-0.745	-0.104	-0.502
	(0.014)	(0.013)	(0.012)	(0.011)	(0.012)
FM3-Γ-linear	-0.411	0.217	-0.577	0.057	-0.334
	(0.014)	(0.013)	(0.012)	(0.011)	(0.012)
ols-linear / N	-0.102	-0.087	-0.336	-0.339	-0.336
	(0.312)	(0.281)	(0.268)	(0.241)	(0.268)
ols-log / N	1.635	1.367	0.170	0.066	0.170
	(0.312)	(0.281)	(0.268)	(0.241)	(0.268)
ols-square root / N	-0.027	-0.019	-0.272	-0.276	-0.272
	(0.312)	(0.281)	(0.268)	(0.241)	(0.268)
ols-linear>0 / N	-0.061	-0.046	-0.376	-0.379	-0.376
	(0.312)	(0.281)	(0.268)	(0.241)	(0.268)
Γ-linear / N	0.134	0.142	-0.176	-0.176	-0.176
	(0.312)	(0.281)	(0.268)	(0.241)	(0.268)
Γ-square / N	0.130	0.150	-0.550	-0.536	-0.550
	(0.312)	(0.281)	(0.251)	(0.241)	(0.268)
FM2-Γ-linear / N	1.389	1.368	0.411	0.415	0.411
	(0.312)	(0.281)	(0.268)	(0.241)	(0.268)
FM3-Γ-linear / N	1.434	1.403	1.057	1.041	1.057
	(0.312)	(0.281)	(0.268)	(0.241)	(0.268)
R^2	0.994	0.995	0.995	0.996	0.995

Table 4
Response Surface Regressions for Mean Absolute Prediction Error

Estimation Sample	Unchanged	Unchanged	Top-coded	Top-coded	Top-coded
Prediction Sample	Unchanged	Trimmed	Unchanged	Trimmed	Top-coded
ols-linear	1.516	1.380	1.484	1.338	1.427
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
ols-log	2.038	1.908	2.001	1.872	1.982
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
ols-square root	1.463	1.314	1.437	1.282	1.378
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
ols-linear>0	1.502	1.364	1.472	1.325	1.414
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Γ-linear	1.431	1.273	1.414	1.251	1.353
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Γ-square	1.448	1.307	1.425	1.277	1.366
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
FM2-Γ-linear	1.409	1.244	1.400	1.233	1.339
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
FM3-Γ-linear	1.431	1.274	1.416	1.254	1.355
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
ols-linear / N	0.156	0.175	0.048	0.055	0.051
	(0.035)	(0.039)	(0.029)	(0.032)	(0.031)
ols-log / N	0.195	0.189	0.044	0.036	0.044
	(0.035)	(0.039)	(0.029)	(0.032)	(0.031)
ols-square root / N	0.036	0.039	-0.005	-0.006	-0.005
	(0.035)	(0.039)	(0.029)	(0.032)	(0.031)
ols-linear>0 / N	0.151	0.169	0.058	0.067	0.061
	(0.035)	(0.039)	(0.029)	(0.032)	(0.031)
Γ-linear / N	0.094	0.108	0.031	0.036	0.03329
	(0.035)	(0.039)	(0.029)	(0.032)	(0.031)
Γ-square / N	0.107	0.118	0.003	-0.002	0.003
	(0.035)	(0.039)	(0.029)	(0.032)	(0.031)
FM2-Γ-linear / N	0.195	0.239	0.096	0.113	0.102
	(0.035)	(0.039)	(0.029)	(0.032)	(0.031)
FM3-Γ-linear / N	0.181	0.223	0.140	0.172	0.149
	(0.035)	(0.039)	(0.029)	(0.032)	(0.031)
R^2	0.999	0.999	0.999	0.999	0.999

Figure 1
Mean Prediction Error as a Function of Sample Size



Unchanged Estimation Sample
Unchanged Prediction Sample

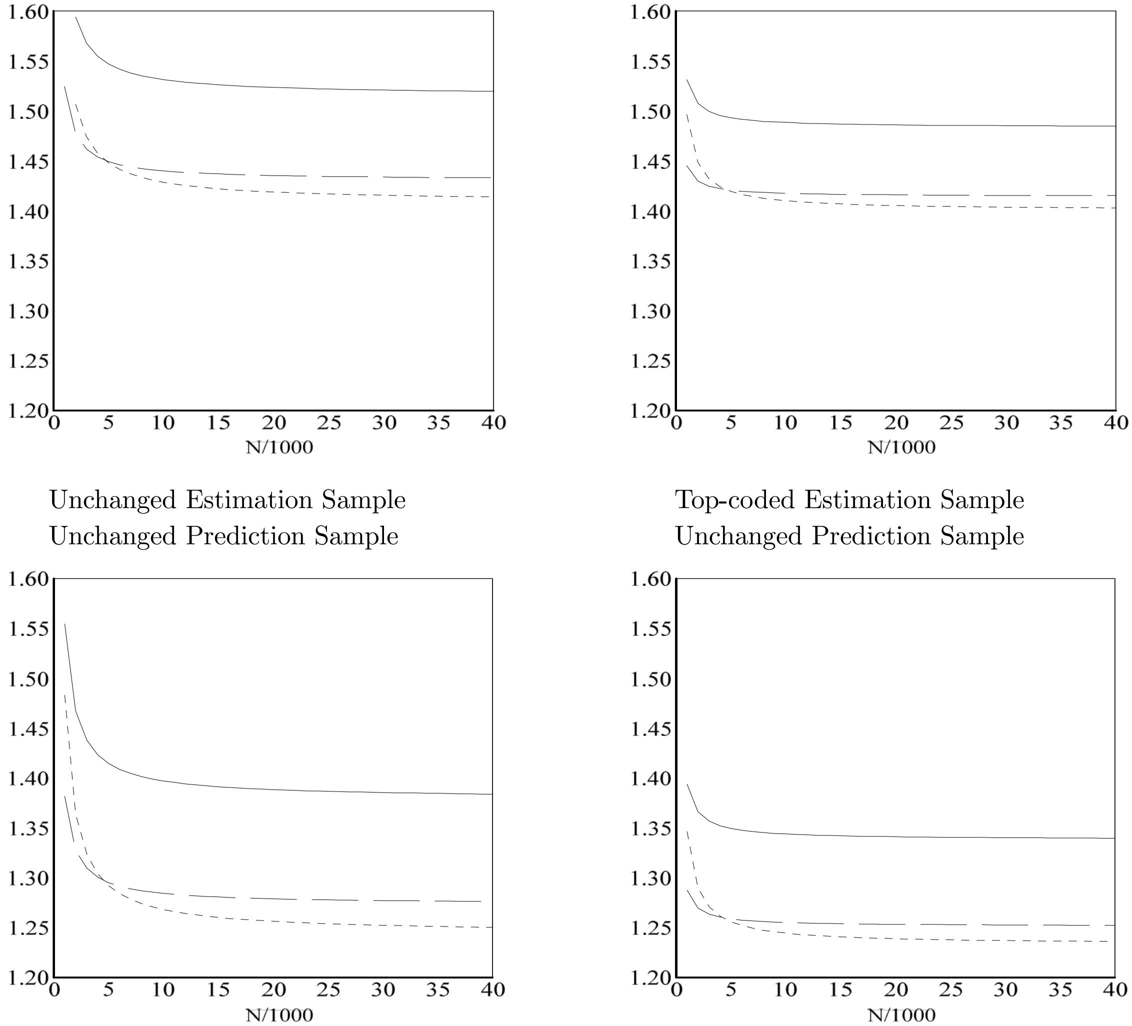
Top-coded Estimation Sample
Unchanged Prediction Sample

Trimmed Prediction Sample
Unchanged Estimation Sample

Top-coded Estimation Sample
Trimmed Prediction Sample

Key: —— ols-linear — — Γ-linear - - FM2-Γ-linear

Figure 2
Mean Absolute Prediction Error as a Function of Sample Size



Unchanged Estimation Sample
Unchanged Prediction Sample

Top-coded Estimation Sample
Unchanged Prediction Sample

Trimmed Prediction Sample
Unchanged Estimation Sample

Top-coded Estimation Sample
Trimmed Prediction Sample

Key: —— ols-linear — — Γ-linear - - FM2-Γ-linear

Appendix: Table A1

Variable	Description	Frequency
HCC020	High Cost Chronic Gastrointestinal	0.010
HCC030	Dementia	0.028
HCC060	High Cost Vascular Disease	0.052
HCC080	Other Urinary System	0.070
HCC100	Minor Symptoms, Signs, Findings	0.323
HCC031	Drug/Alcohol Dependence/Psychoses	0.065
HCC051	Other Acute Ischemic Heart Disease	0.011
HCC091	Chronic Ulcer of Skin	0.017
HCC022	Moderate Cost Gastrointestinal	0.048
HCC032	Psychosis/Higher Cost Mental	0.088
HCC042	High Cost Neurological	0.022
HCC013	Diabetes with Chronic Complications	0.037
HCC023	Low Cost Gastrointestinal	0.177
HCC033	Depression/Moderate Cost Mental	0.070
HCC043	Moderate Cost Neurological	0.049
HCC053	Valvular and Rheumatic Heart Diseas	0.022
HCC063	Other Circulatory Disease	0.024
HCC113	Elective/Aftercare	0.129
HCC004	Other Infectious Disease	0.132
HCC014	Diabetes with Acute Complications	0.018
HCC044	Low Cost Neurological	0.042
HCC064	Chronic Obstructive Pulmonary Disea	0.118
HCC015	Diabetes with No or Unspecified Com	0.128
HCC025	Rheumatoid Arthritis/Connective Tis	0.018
HCC075	Low Cost Ear, Nose, and Throat	0.184
HCC006	High Cost Cancer	0.011
HCC116	Rehabilitation	0.036
HCC007	Moderate Cost Cancer	0.012
HCC017	Moderate Cost Endo/Metab/Fluid-Elec	0.026
HCC067	Low Cost Pneumonia	0.016
HCC097	Other Injuries and Poisonings	0.111
HCC008	Low Cost Cancers/Tumors	0.050
HCC028	Blood/Immune Disorders	0.013
HCC048	Congestive Heart Failure	0.059
HCC058	High Cost Cerebrovascular Disease	0.012
HCC078	Renal Failure	0.020
HCC098	Complications of Care	0.017
HCC118	History of Disease	0.063
HCC029	Iron Deficiency and Other Anemias	0.048
HCC049	Heart Arrhythmia	0.043
HCC059	Low Cost Cerebrovascular Disease	0.043
HCC099	Major Symptoms	0.156